METHOD AND DEVICE FOR EXTRACTING KNOWLEDGE FROM ENORMOUS DOCUMENT DATA AND MEDIUM

Publication number: JP2001084250

Publication date: 2001-03-30

Inventor: MATSUZAWA YASUSHI; FUKUDA TSUYOSHI; NASUKAWA TETSUYA; NAGANO TORU; MOROHASHI MASAYUKI

Applicant: IBM

Classification:

- International: ***G06N5/04; G06F9/44; G06F17/27; G06F17/28; G06F17/30;*** **G06N5/00; G06F9/44; G06F17/27; G06F17/28; G06F17/30;** (IPC1-7): G06F17/30; G06F9/44; G06F17/27

- European:

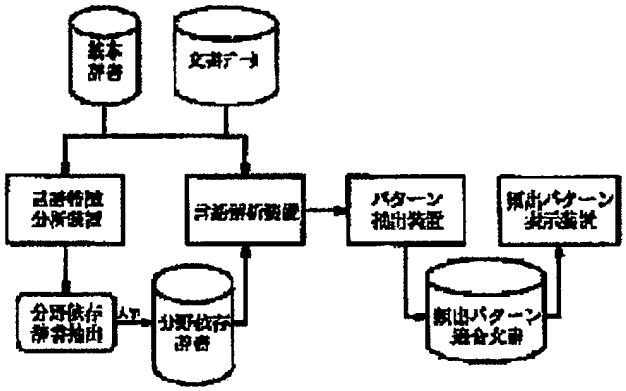
Application number: JP19990239674 19990826

Priority number(s): JP19990239674 19990826

Report a data error here

Abstract of **JP2001084250**

PROBLEM TO BE SOLVED: To automatically extract a document satisfying a pattern from enormous amount of documents, to extract useful knowledge and to reduce time required for a response by generating a field-dependent dictionary from document data, generating a syntax tree considering modification, by means of a language analysis device and extracting/outputting a frequentlyappearing pattern by means of a pattern extraction device. SOLUTION: A language feature analysis device generates an analysis- dependent dictionary. A language analysis device needs to prepare a field- dependent dictionary for requiring an attribute adjusted to data to be analyzed. A word having the specified attribute is to be generated by each field. The language feature analysis device checks the word from actual data and registers it in the field-dependent dictionary. A pattern extraction device obtains a pattern, which frequently appears by using document data which is structure-analyzed by the device and takes out an original document having a syntax which is matched with the pattern. A frequently-appearing pattern device displays the document, having the detected frequently-appearing pattern and a syntax tree matched with it.



Data supplied from the *esp@cenet* database - Worldwide

(19)日本国特許庁（ＪＰ）　(12)公開特許公報（A）　(11)特許出願公開番号
特開2001－84250
(P2001－84250A)
(43)公開日　平成13年３月30日(2001.3.30)

(51)Int.Cl.'		識別記号	FI		テーマコード'(参考)	
G06F	17/30		G06F	15/40	370A	5B075
	9/44	550		9/44	550N	5B091
	17/27			15/38	E	
	17/28				U	
				15/40	380A	

審査請求 有　請求項の数9　OL　(全９頁)　最終頁に続く

(21)出願番号　特願平11－239674

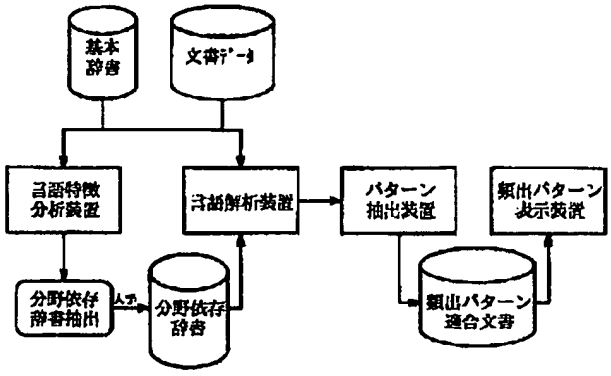
(22)出願日　平成11年８月26日(1999.8.26)

(71)出願人　390009531
インターナショナル・ビジネス・マシーンズ・コーポレーション
INTERNATIONAL BUSINESS MASCHINES CORPORATION
アメリカ合衆国10504、ニューヨーク州
アーモンク　(番地なし)
(74)代理人　100086243
弁理士　坂口　博　(外４名)

最終頁に続く

(54)【発明の名称】　膨大な文書データからの知識抽出方法、その装置及び媒体

(57)【要約】
【課題】　キーワードの出現順序に着目したデータマイニング手法では、同一のものとして扱われてしまい、より細かく分類して知識抽出できない文書があった。また、順序を無視した共起関係のみを扱うマイニング手法でも誤って全く異なる意味の知識を同一のものとして知識抽出してしまう場合があった。更に抽出された知識が人間にとって理解が困難であった。
【解決手段】　本発明は、言語特徴分析装置によって言語解析装置の精度向上のために文書データから分野依存辞書を作成し、言語解析装置によって共起関係と係り受けを考慮した構文木を作成し、パターン抽出装置が、この構文木を用いることによって適切に頻出パターン（即ち、知識）を抽出・出力する。



1

【特許請求の範囲】

【請求項１】　大量の文書データからの知識抽出を行う方法において、形態素解析技術により１つの文書から単語を切り出し単語間にある係り受けの関係を推定し係り受け関係から構文木を構築するステップと、構築された構文木の中で多くの構文木に含まれている頻出パターンをパターンの制約に基づいて発見するステップと、発見された頻出パターンへの代入にマッチする文書を検索するステップと、を含むことを特徴とする知識抽出方法。

【請求項２】　前述した構文木を構築するステップにおいて、線形リストを構築し、構築した線形リストをも使用して頻出パターンを発見する、請求項１に記載の知識抽出方法。

【請求項３】　前述した頻出パターンを発見するステップにおいて、前記線形リストを用いて、探索範囲と単語とラベルとの組み合わせを正規表現を用いて記述されたパターンを探索して知識を抽出することを特徴とする、請求項２に記載の知識抽出方法。

【請求項４】　大量の文書データからの知識抽出を行う装置において、基本辞書に含まれない語彙を分野依存辞書に登録する言語特徴分析装置、自然言語解析を行う言語解析装置、パターンの制約に基づいて特定パターンに適合するデータを発見するパターン抽出装置及び、抽出した頻出パターンを表示する頻出パターン表示装置を具備し、文書データから一般分野を対象とする基本辞書と、文節生成処理用の生成規則と、構文木生成用の生成規則と、分野依存辞書と、を参照して知識抽出を行う、ことを特徴とする知識抽出装置。

【請求項５】　前記言語特徴分析装置は、形態素解析用辞書を用いて入力文書を品詞付き単語列に分割し、分野依存辞書を用いて既に登録されている語を単語列から削除し、残った語に対して出現頻度を計算し、頻度の多い順序に並び替え、分野依存辞書に追加登録する手段を含む、ことを特徴とする請求項４に記載の知識抽出装置。

【請求項６】　前記言語解析装置は、形態素解析装置、文節生成装置、辞書適用装置、及び係り受け解析装置を含み、文節生成規則及び構文木生成規則に応じて、距離、係り受け、及びラベルを考慮して、線形リスト及び構文木の形態で構文解析データを生成する手段を含み、前記形態素解析装置は、入力文書を形態素解析を用いて各単語に分割し、品詞または属性を含むラベルを付加し、同義語辞書を用いて表現を統一させる手段を含む、ことを特徴とする請求項４に記載の知識抽出装置。

【請求項７】　前記パターン抽出装置は、頻出パターン抽出装置と特定パターン適合文書抽出装置を含み、前記頻出パターン抽出装置は、構文解析データを用いて、単語と、単語の位置関係と、ラベルとの組み合わせに基づいて、共起関係を調べ、頻出するパターンを抽出する手段を含み、前記特定パターン適合文書抽出装置は、構文解析データが特定のパターンを構築する単語、属性を含むか否か、各文節間に係り受けの関係があるか否かを検査することによって、頻出パターンに一致する文書を抽出し、これを出力する手段を含む、ことを特徴とする請求項４に記載の知識抽出装置。

2

【請求項８】　前記頻出パターン表示装置は、前記パターン抽出装置によって発見された頻出パターンとこれに合致する構文木を持つ文書の表示手段を含む、ことを特徴とする請求項４に記載の知識抽出装置。

【請求項９】　大量の文書データからの知識抽出を行うプログラムにおいて、形態素解析技術により１つの文書から単語を切り出し単語間にある係り受けの関係を推定し係り受け関係から構文木を構築するステップと、構築された構文木の中で多くの構文木に含まれている頻出パターンをパターンの制約に基づいて発見するステップと、発見された頻出パターンへの代入にマッチする文書を検索するステップと、をコンピュータに実行させるためのプログラムを記録したコンピュータ読み取り可能な媒体。

【発明の詳細な説明】

【０００１】

【発明の属する技術分野】本発明は、特定の分野を対象とした大量の文書から知識抽出を行うために、自動的にパターンを抽出する技術に関するものであり、特に、抽出された特定のパターンを満足する文書を大量文書の中から抽出することによって、有用な知識抽出を行う膨大な文書データからの知識抽出方法、その装置及び媒体に関する技術である。

【０００２】

【従来の技術】計算機及びネットワーク環境の発達と普及により、膨大なデータが電子化されて蓄積され、オンラインで参照可能となっている。このデータを有効利用すべく、データマイニングの技術が盛んに研究開発されてきた。しかし、従来のデータマインニング技術で対象としているのは、数値を中心とした集計可能な定型データのみである。しかし、大抵のデータにはテキスト部分（即ち文書データ）が含まれており文書データは基本的に定型ではないため、数値を中心とした定型データと異なり集計が困難である。従って、文書データについては、基本的には１つ１つ目を通す必要があるため、非常に手間がかかってしまう。即ち、人手で分析できる文書データの量には限度があり、せっかく蓄積された膨大な文書データを持て余してしまうという問題が生じている。このような、非定型のテキスト文書から知識を抽出する技術は、「テキストマイニング」と呼ばれ最近注目を浴びている。このテキストマイニングは、コールセンターの記録、アンケート結果の集計等での利用だけでなく、特許関係の文書、営業報告書等あらゆる非定型な文書の分析に応用可能なことから最も期待されている技術である。

【０００３】大量の文書の内容を分析する手段として、

類似内容を持つ文書を見つけてカテゴリごとに分類する方法がある。例えば、現在ウェブの検索サイト等において使われている方法として、予めカテゴリを用意しておき、人が文書を読みその文書が該当するカテゴリを判断し分類するというものがある。また、特定のキーワードを含む文書はあるカテゴリに属するというルールに基づいて、この作業を自動的に行うことも可能である。例えば、「ABS」、「エアバッグ」というキーワードを含む文書があれば車というカテゴリに属すると判断できる。これは大量文書の大分類には適するが、より細かい分類を行うことは困難である。

【０００４】例えば、コールセンター業務においては、顧客からの電話内容にはどのような要件が多いのかを分析することによって、コールセンター業務を改善したいという要求がある。電話を記録した内容を人手によって大雑把に分類し、分類した結果から注意深く文書を読み、ほぼ同一内容の文書を集計する作業で、この要求は達成できる。しかし、毎月、何万件という問い合わせを受けるコールセンターの場合、人手で、これを行うのは非常に労力がかかり、現実には困難である。また、蓄積された文書は、特定分野を対象とした文書であり、カテゴリを非常に細かく分ける必要があるが、内容を予測して事前にカテゴリを用意するのも非常に困難である。例えば、簡単な「車」というカテゴリではなく、更に細かく「エンジンの異音の発生」等と細かく分類することが要求される。このような細かい分類では、分類する人は文書の内容を更に良く吟味して分類作業をしなければならず、その作業量は膨大となる。また、カテゴリの判断基準が人によって異なったり、同一人物でもその都度違う判断をする可能性があり、客観的なデータを得ることが難しい。

【０００５】近年、計算機を用いた文書の分類手法（文書のクラスタリング）が開発されているが、この手法は文書中に出現するキーワードから大雑把な分類を行うものである。しかし、特定分野を対象とする場合には、より細かな分類が必要であり、従来の手法では対処できない。また、クラスタリングの結果、どんな内容の文書が１つのクラスタに集められたのかは、その文書を人が読まなければ理解できないという問題点がある。

【０００６】上述のように、大量の文書から語をキーワードとして切り出し、共起する単語のペアを取り出す従来技術が、データマイニングにおける「相関ルールの抽出技術」と呼ばれるものである。しかし、この手法では以下の問題点がある。長い文書において始めに現れる語と最後に現れる語との間には関連性が無い場合があるが、これを共起するものとしてカウントしたり、語の係り受けの関係が無視されているために、例えば「ＡがＢするとＣがＤした」と「ＡがＤするとＣがＢした」では意味が異なるが、共起関係だけを見ると、これら２つの文書を同じものとして処理してしまう。従って、同一内容の文書抽出が正しく行われない場合が多い。

【０００７】上述のような、不都合を解決するためには、特定の単語が特定の順番で出現するものだけを抽出する方法が考えられる。これがデータマイニングにおける「時系列パターン抽出技術」と呼ばれるものである。例えば、単語Ａ、単語Ｂ、単語Ｃ、単語Ｄという順序で単語が出現する文書だけを抽出することができる。しかし、このルールでは「ＡがＢするとＣがＤした」という文書の場合は抽出できるが、「Ｃは、ＡがＢすると、Ｄした」という文書は、文書の意味は同じだが、単語の順番が異なっているため抽出できないという問題がある。即ち、同一内容の文書を抽出するためには、単語の共起関係、出現順序だけでなく、単語間の係り受けの関係にも着目する必要がある。

【０００８】

【発明が解決しようとする課題】上述のように、本発明では、大量の文書から特定のパターンを抽出すること、また、そのパターンを満足する文書を自動的に抽出することにより、有用な知識抽出を実現する膨大な文書データからの知識抽出方法、その装置及び媒体を提供するものである。

【０００９】

【課題を解決するための手段】本発明は、大量の文書データからの知識抽出方法を対象とする。この知識抽出方法において、形態素解析技術により１つの文書から単語を切り出し単語間にある係り受けの関係を推定し係り受け関係から構文木を構築するステップと、構築された構文木の中で多くの構文木に含まれている頻出パターンを与えられたパターンの制約に基づいて発見するステップと、発見された頻出パターンへの代入にマッチする文書を検索するステップと、を含むものである。また、本発明は、上記方法のステップをコンピュータに実行させるためのプログラムを記録したコンピュータ読み取り可能な媒体をも含むものである。

【００１０】更に、本発明は、大量の文書データからの知識抽出装置を対象とする。この知識抽出装置において、基本辞書に含まれない語彙を分野依存辞書に登録する言語特徴分析装置、自然言語解析を行う言語解析装置、特定パターンに適合するデータを発見するパターン抽出装置及び、抽出した頻出パターンを表示する頻出パターン表示装置を具備し、文書データから一般分野を対象とする基本辞書と、文節生成処理用の生成規則と、構文木生成用の生成規則と、分野依存辞書とを含む。上記構成において、大量文書からの知識抽出を好適に実施できる。

【００１１】

【発明の実施の形態】言語特徴分析装置によって言語解析装置の精度向上のために文書データから分野依存辞書を作成し、言語解析装置によって係り受けを考慮した構文木を作成し、パターン抽出装置によって頻出パターン

（即ち、知識）を抽出・出力する。以下、装置の形態で発明を説明するが、本発明には、方法、プログラム媒体も含まれることは言うまでもない。具体的な機能としては、

１．　形態素解析技術により、１つの文書から単語を切り出し、単語間における係り受けの関係を推定し、係り受け関係から構文木を構築する機能、

２．　大量の文書から構築された大量の構文木の中で、与えられたパターンの制約に基づいて、多くの構文木に含まれているような頻出パターンを発見する機能、

３．　発見された頻出パターンの構文木を有する文書を出力する機能、等である。

【００１２】図１は、文章から形態素を切り出し、係り受け関係を抽出し、その係り受け関係から構文木を生成する過程の概略図を示している。図１の文章”ＡがＢすると、ＣがＤする”から、形態素解析、係り受け関係の抽出を行った結果、「Ａ」が「Ｂする」、「Ｃ」が「Ｄする」、「Ｂする」と「Ｄする」という２項関係が抽出される。ここで２つの単語間の係り受け関係から矢印の向きが決まる。これらの関係から、図の構文木が生成される。構文木は有向グラフ（接点を結ぶ枝に向きが有るグラフ）として表現される。有向グラフ上の節点（ノードと呼ぶ）には、形態素解析で切り出した単語をラベルとして付与する（図中では、Ａ～Ｄと略記）。２つのノード間を結ぶ枝（アークと呼ぶ）には向きがある。アークの向きは、前述のように、単語間の係り受け関係により決まる。図２（ａ）のように、ここで、パターンとは、構文木中に存在するノードとその位置関係を示す。ノード、即ち単語の個数は任意である。ここで、各単語に対して制約を与えることができる（例えば、動詞、専門用語であること等）。位置関係は、一定のものに制約しても良いが、単語が少数であれば可能性のある全ての位置関係であっても良い。パターンの例を示す。いま、１つの構文木中に、２つの単語Ａ、Ｂがあったとき、Ａというラベルを持つノードからＢというラベルを持つノードに構文木中の有向グラフを辿ることで、到達することができ、更に図２（ｂ）のように、それがある距離内であるとき、これをＡ－＊→Ｂと記述し、これをパターンとすることができる。更に、同様にして、他の単語Ｃ、Ｄがあって、同時にＡ－＊→Ｂ、Ａ－＊→Ｃ－＊→Ｄの関係が成り立っているとき、これを４つの単語とその位置関係からなるパターンとする。また、このパターンに対しても制約を与えることができる。例えば、上記Ａに対して動詞である、専門用語である等の制約である。頻出パターンの発見とは、このように複数の単語とその位置関係を表わすパターンのうち頻出するものを発見することである。

【００１３】文書が日本語等の場合は、構文木だけでなく線形リストを構築することもできる。線形リストに対しても、同様に与えられたパターンの発見をすることができ、この場合は処理が高速化される。

【００１４】共起関係については、一般的に文章中の語句と語句との距離が大きくなるほど、その語句と語句との関連性が小さくなることが多いため、距離（例えば、構文木において、あるノードからあるノードまでに経過する枝の数（アーク数））という概念を導入する。例えば距離＝３と定義する場合は、距離が４以上あるような、語句と語句が離れているノード間を共起関係が無いものとして取り扱う等である。この距離は、対象の文書に応じて適切な値を設定する。図３は本発明の全体構成を示す図である。また、図４は本発明の処理の流れを示すフローチャートである。図５は言語解析装置の詳細を示すものであり、本装置によって構造解析された文書データを用いて、パターン抽出装置は頻出するパターンを求め、そのパターンと合致する構文を持つ元の文書を取り出す。頻出パターン表示装置は、発見された頻出パターンとそれに合致する構文木を持つ文書を表示する。ここで、本発明を構成する１．言語特徴分析装置、２．言語解析装置、３．パターン抽出装置及び、４．頻出パターン表示装置について説明する。

【００１５】１．言語特徴分析装置について

言語特徴分析装置は、言語解析装置の精度を向上させるために分野依存辞書の作成を行う。これは、一般的な辞書に含まれていない特定分野のための語彙を追加し、その語彙の属性について記述する。また、分野によって意味や属性が異なる語彙について分野依存辞書を作成する。言語解析装置は、分析するデータに合った属性を必要とするため、分野依存辞書（例えば「装置（１９）」を「装置（Ｈ／Ｗ）」に書きかえるための辞書）を用意する必要がある。「装置」や「良－い」といった一般語については、最初に用意したものをどのデータに対しても利用できるが、製品名のような特定の属性を持つ語などは、分野ごとに作成しなければならない。これを、実際のデータから調べて分野依存辞書に登録するのが言語特徴分析装置であり、以下の手順で登録を行う。

【００１６】Ａ．従来技術である形態素解析装置と基本辞書を用いて文を品詞付き単語列に分割する。

Ｂ．分野依存辞書に既に登録済みのものは単語列から削除する。

Ｃ．単語の出現頻度を計算し、単語列を出現頻度の多い順に並べ替える。

Ｄ．この単語列の中から、予め設定した属性に該当する言葉を見つけて分野依存辞書に追加登録を行う。ここで、分野依存辞書中のエントリーの構造を品詞付き単語列→品詞または属性付き単語列という形にすれば、たとえ形態素解析装置が誤った単語分割や誤った品詞付与をしても必要な単語と属性を取り出すことができる。

【００１７】２．言語解析装置について

言語解析装置は、形態素解析装置、文節生成装置、辞書適用装置、及び係り受け解析装置を含むものであり、以

下各々について説明する。

（１）形態素解析装置

入力された文に対して従来技術である形態素解析を行うことによって単語ｔに分割した後、基本辞書を用いて単語列に対してラベル１（品詞あるいは属性名に相当する名前）を付加する。また単語間の距離ｄを重みとして付加する。以下、形態素 w＝[ｔ，１，ｄ]の組とする。また同義語辞書を用いて、表現のゆれや同義語を１つの統一された表記に変更する。

【００１８】（２）文節生成装置

文（あるいは特定の文脈）に各語句が出現する順番にw_1，w_2，，…．w_nとすると、w_1から順に生成規則に従って文節を決定する。w_nが付属語である場合や、明らかに文節が切れると判断できるところで文節を区切る。w_kで文節を区切られた場合、次の文節はw_k+1から始まり、これを文末になるまで行う。各文節を自立語と付属語の組合わせにし、これを構文木のノード及びノードからのアークとする。また、「反、非」等の接頭語、「ない」等の助動詞がある語句の場合は、ラベルの符号を反転させる。

【００１９】（３）辞書適用装置

分野依存辞書によって、単語列中の単語及びラベルを書き換える。対応する属性名が無い場合は、品詞がそのままラベルとして残る。各ノードには単語の他に、品詞等の情報、アークには助詞の情報等が付加される。

【００２０】例えば、「装置が良くない訳ではない」という文章からは、下記のようなものが出力される。ここで用いた形態素解析装置においては常に重みは１になり、重みｄの表示を省略する。また番号は品詞を示す。例えば、19…名詞、75…格助詞「が」、17…形容詞の語幹、42…形容詞連用形活用語尾、等である。句点（。）のｄを∞にすること等は、簡単ではあるが効果的な重み付けの方法である。

（１）．形態素解析装置からの出力：［装置，19］
［が，75］［良－い，17］［く，42］［な－い，51］［い，43］［訳，94］［で，56］［は，85］［な－い，51］［い，43］

（２）．文節生成装置からの出力：（［装置，19］［が，75］）（［良－い，17］［く，42］［な－い，51］［い，43］）（［訳，94］［で，56］［は，85］）（［な－い，51］［い，43］）小括弧で区切られているのが文節である。

（３）．辞書適用装置からの出力：（［装置，H/W］［が，75］）（［良－い，評価］［く，42］［な－い，51］［い，43］）（［訳，94］［で，56］［は，85］）（［な－い，51］［い，43］）

このように、入力文章から文節毎に分解されて、線形リストの構文構造データが作成される。更に、後述する文節間の係り受け関係の分析をすることで、有向グラフの構文構造データを作成することができる。

【００２１】（４）係り受け生成装置

文法規則は、係り受け元のノードの自立語（R_{sd}）、付属語（R_{si}）、係り受け先の自立語（R_{dd}）と付属語（R_{di}）、及び係り受けの性質(T)、の組み合わせ｛R_{sd}，R_{si}，R_{dd}，R_{di}，T｝から構成される。この文法規則を係り受け元のノードN_nと係り受け先のノードN_m（n m）に適用し、文法規則に合致した場合N_nとN_mに係り受けの関係があると判断し、N_nからN_mに対して係り受けの関係をつける。文法規則に合致すれば、係り受けは幾つでも持つことができる。また付属語及び係り受けの性質からアークに重みを付けることもできる。抽出した係り受けの関係をアークとし、辞書適用装置で抽出した情報を各ノードに付加することによって、構文木を作成する。

【００２２】３．パターン抽出装置について

パターン抽出装置は、頻出パターン抽出装置と特定パターン適合文書抽出装置を含むものであり、以下各々について説明する。

【００２３】（１）頻出パターン抽出装置

ここでは、１つのパターンとして、４つの単語（仮にVa,Vb,Na,Nbとする）とその位置関係としてVa－＊→Vb－＊→Nb、Va－＊→Naを考える。またVa、Vbは動詞であること、Na、Nbは名詞であることを制約として与える。このようなパターンが与えられると頻出パターン抽出装置は、各構文木に含まれる単語で、VaとNa、VbとNb、VaとVbという係り受けの関係を持ち、かつVa Vbが動詞、Na Nbが名詞であるような単語の組（Va －Vb－ Na－ Nb）を探し、これを集計していく。

【００２４】実現するための一例として具体的には、

（１）Ａ．構文木を解析し、動詞ノードを見つけ、そのノードから近距離に存在する動詞ノードについて調べ、動詞と動詞の係り受けの関係にある動詞－動詞のペアを求める。経路が複数ある場合は、距離が最短となるルートでの距離を集計の対象とする。例えば、ノードVaから有向グラフを辿っていき、一定の距離内にあるノードVbが存在すればノードVaとノードVbのペアが対象となる。これを構文木上の全ての動詞ノードに対して行う。例えば、ここでVa－Vb、Vb－Vcが発見されたこととする。

Ｂ．Ａと同様に、構文木を解析し、動詞ノードから近距離に存在する名詞ノードについて調べ、係り受けの関係にある動詞－名詞のペアを求める。例えば、ここでVa－Na、Vb－Nb、Vc－Ncのペアが発見されたとする。

Ｃ．Ａで求めた動詞－動詞の係り受けのペアと、Ｂで求めた動詞－名詞の係り受けのペアから4つの語からなる組を求める。例えば、ＡでVa－Vbが発見されて、かつＢでVa－Na、Vb－Nbが発見されれば、図７のように、この４つの語からなる組（Va－Na－Vb－Nb）は集計対象となる。同様に（Va－Nb－Vc－Nc）も集計対象となる。

【００２５】（２）全ての文書（構文木）に対して、上記Ａ、Ｂ、Ｃを行い、最終的に集計された４つの語から

なる組の中から、頻出した組み合わせを出力する。

（３）要素数の多い頻出パターンを抽出する場合を考える。パターンとして６つの単語（Va, Vb, Vc, Na, Nb, Nc）からなり、Va－＊→Vb－＊→Vc－＊→Nc、Va－＊→Na、Vb－＊→Nbという位置関係を考える。また、Va、Vb、Vcは動詞であること、Na、Nb、Ncは名詞であることを制約として与える。このようなパターンが与えられた時には、同様にして、Ａで求めた動詞－動詞のペアの中に、Va－Vb, Vb－Vcというペア（VaはVbに、VbはVcにそれぞれ係り受けの関係がある）が存在するか調べ、Ｂで求めた動詞－名詞の係り受けのペアを用いて、図８のように６つの語からなる組を抽出する。

【００２６】（２）特定パターン適合文書抽出装置

大量文書の中から、頻出パターンを満足する文書を抽出し、これを出力する。これは、構文解析データ（構文木データ）に対して、特定のパターンを構築する単語や属性を全て含んでいるか、含んでいる場合には、それぞれ

$$P=<p_1, p_2, \cdots, p_n>$$

で表わすことができる。各p_i（$i=1,2,\cdots,n$）は、単語ｔと品詞または属性を表わすラベル１の組であり、Ｐはこの$p_1, p_2, \cdots, p_n$を順に並べたものである。このとき、p_iと次に続くp_{i+1}は、（１）で指定した係り受けの探索範囲以内に存在しなければならない。また、パターンは、正規表現を用いて記述することもできる。このパターンＰに一致するものを文章[ｔ，１，ｄ]＊の中から探索し、これに一致する線形リストの部分集合の重み付き距離

$d=\Sigma(d_1, \ldots, d_n)$

（$d_1, \ldots, d_n$はパターンにマッチする最初から最後までのワードの重み付き距離）が最少となるものを選び出す。

【００２８】（３）探索範囲と探索パターンを与えられて、入力の単語列[ｔ，１，ｄ]＊（単語は名前ｔ、属性名１、右隣の単語との距離ｄという要素からなる）からパターンに合致する単語の組を取り出したものが、抽出情報である。例えば、「装置が良くない訳ではない」という文を例にとると、この文から構築された線形リスト（

【００２０】参照）から、パターン

Ｐ＝＜［＊，Ｈ／Ｗ］＞

により属性名「Ｈ／Ｗ」にマッチする要素［装置，Ｈ／Ｗ］（距離は省略）を取り出すことができる。

Ｐ＝＜［＊，Ｈ／Ｗ｜Ｓ／Ｗ］，［＊，評価］＞

により、テキスト中から複合属性［Ｈ／Ｗ］－［評価］または［Ｓ／Ｗ］－［評価］にマッチする要素の組を探し、この例では［装置，Ｈ／Ｗ］－［良い，評価］を取り出すことができる。

【００２９】４．頻出パターン表示装置について

パターン抽出装置によって発見された頻出パターンとそれにマッチする構文木を有する文書を表示する。

の単語間に係り受けの関係があるのか否かを調べることで実現できる。

【００２７】（３）線形リストからのパターン抽出

言語解析装置において、係り受け解析装置にかけるデータとして、線形リストの構造を持つ構文解析データが構築されており、このデータからも以下のようにパターンを抽出することが可能である。

（１）重み付きの距離を含んだ形態素（線形リストの要素）ｗの列ｗ＊に対し、係り受けの探索範囲を０～∞で設定する。ｗは単語ｔ，品詞または属性を表わすラベル１、右隣の単語との重み付き距離ｄの組である（ｗ＝［ｔ，１，ｄ］）。この時、探索範囲の値が０というのは、探索を開始する場所の単語のみを探すことを意味し、１ならば前後の単語も係り受けの探索候補とすることを意味する。

（２）探索パターンは

$$p_1, \cdots, p_n \in \{[t, 1]\}$$

【００３０】本手法を実際のコールセンター業務で作成された９万文のコールデータを処理して、その有効性を確認した。以下に実施例の１具体例を示す。始めに個々の文書から従来技術である形態素解析を行い、係り受け解析装置によって構文木を構築する。例として簡単な文章「電源を入れるとフロッピーディスクを要求する絵が出る。」を用いることとする。この文章からは図９のような構文木（有向グラフ）が構築される。このグラフ中で、有向のアークは語句の係り受けの関係を表わしている。また、ノード（各語句）の右肩にある四角は、その語が動詞であるか名詞であるかを示す（Ｎは名詞、Ｖは動詞を示す）。

【００３１】この構文木を作成するための文法規則は８５個であり、あるノードの語句が動詞の連体形であれば、そのノード以降に現れる名詞に対して係り受けを行うというような簡単なものである。この例では、アークの重みは全て等しく１とする。有向グラフにおいて、あるノードからあるノードまでに経過した枝の数（アーク数）を距離と定義する。例えば、「電源」と「要求する」では２つのアークを経由することで到達できるので、距離＝２となる。複数の経路が存在する場合は最短の経路で計算する。また、抽出する知識としては、ここでは距離が３以内のものだけを考えることとする。このように距離をある程度短くすることで、単語間の関連性が無いと推定される係り受けを排除することが可能となる。上記の構文木から動詞－名詞の係り受けを求めると、「出る」－「絵」、「要求する」－「フロッピーディスク」、「入れる」－「電源」等の近距離に存在する語句のペアを取り出すことができる。

【００３２】更に、動詞－動詞の係り受けにペアを求めると、「要求する」－「入れる」、「出る」－「入れる」を求めることができる。求めた動詞－動詞、動詞－

名詞の各ペアから、Ｖ１－Ｖ２、Ｖ１－Ｎ１、Ｖ２－Ｎ２の係り受けの関係になっているものを求めると「電源」「入れる」「フロッピーディスク」「要求する」や「フロッピーディスク」「要求する」「絵」「出る」という４つの語からなる組を抽出することができる。また、「電源」「入れる」「フロッピーディスク」「要求する」「絵」「出る」という６つの語からなる組も抽出できる。このように抽出した４つの語からなる組と６つの語からなる組を集計することで、大量文書の中から同じ単語を同じ係り受けの構造の中で用いる文書について集計することができる。

【００３３】「名詞２」－「動詞２」、「名詞１」－「動詞１」、「動詞１」－「動詞２」という構成の４つの語からなる組（即ち知識）を、実際のコールセンターのコール記録文書から抽出してみる。「増設Ｈ／Ｗ」－「外す」、「BIOS」－「戻す」という４つの語からなる知識を抽出することができた。この知識の抽出元となった文章は以下のものである。「増設H/Wを外してBIOSの復元、FDISKで区画の切り直しリカバリーCDで出荷時に戻してください」、「増設H/Wを外してBIOSの復元、リカバリーCDで出荷時に戻していってもISDNカードが使えない」、「増設H/Wを全て外してBIOSをＦ５で工場設定値に戻してもレジューム機能の項目が復活できず、BIOS、H/Wの不具合と考えサービスセンターにて調査が必要と判断」等である。

【００３４】その他に「ファイル」－「見つからない」、「メッセージ」－「出る」という４つの語からなる知識も抽出することができた。この知識の抽出元となった文章は以下のものである。「プログラムファイルエラーのファイルが見つからないとメッセージが出る」、「“または必要なファイルが見つかりません”のメッセージが出るようになったのでメッセージを消したい」、「Xで¥INSTALLと入力しても“ファイルが見つかりません”といった旨のエラーメッセージが出てしまいインストールできない」等である。

【００３５】また、他に「PC」－「表示する」、「OS」－「戻る」、「方法」－「分からない」という６つの語からなる知識も抽出することができた。この知識の抽出元となった文章は以下のものである。「PCの機種A、黒い画面に白い文字が表示されていて、××モードからOSに戻る方法が分からない」、「PCの機種A、ゲーム選択後、コマンドプロンプトが表示され、OSに戻る方法が分からない」、「PCの機種A、日本語DOSゲームアイコン選択後、黒い画面に白い文字で“Cで¥OS”と表示され、OSに戻る方法が分からない」等である。

【００３６】更に、他に「電源」－「入れる」、「フロッピーディスク」－「要求する」、「絵」－「出る」という６つの語からなる知識も抽出することができた。この知識の抽出元となった文章は以下のものである。「電源を入れるとフロッピーディスクを要求する絵が出る」、「ネットワークの設定を確認しようとしたが電源を入れるとフロッピーディスクを要求する絵が出てOS起動できない」、「電源を入れるとフロッピーディスクを要求する絵が出てくる、BIOSでハードディスクは認識している」等である。

【００３７】更に、他に「インターネット」－「接続する」、「発信音」－「聞こえない」、「メッセージ」－「出る」という６つの語からなる知識も抽出することができた。この知識の抽出元となった文章は「機種Aのインターネットに接続しようとすると“発信音が聞こえません”とメッセージが出て繋がらない」、「インターネットに接続しようとすると“発信音が聞こえない”というメッセージが出て接続できない」、「機種Aのインターネットでプロバイダーに接続しようとすると“発信音が聞こえません”とメッセージが出る」等である。

【００３８】本発明による知識抽出（頻出パターン発見）方法のメリットとしては、

（１） 従来法であるキーワードだけを使った共起関係や順序関係のデータマイニングの適用では得ることができなかったパターンを抽出することができる。また従来技術では、誤って見つけてしまうパターンを見つけない。

（２） 抽出された知識（頻出パターン）が人間にとってわかりやすく、視認性に優れる。

（３） 線形リストを併用することで、処理を高速化できる。等がある。

【００３９】

【発明の効果】本発明によって、従来のデータマイニング手法では発見できなかったりまたは誤って発見していた知識を、より適切に誤ることなく知識抽出できるようになった。また、抽出した知識も視認性に優れ、人間にとって理解しやすいものとなった。例えば、企業のコールセンター等では、大量の文書に出現するほぼ同一内容の文書を発見し、出現数の多い内容について調べることで、顧客からの問い合わせの多い内容に対してＦＡＱの作成を行ったり、企業のホームページに掲載することで、問い合わせ件数の低減をすることができたり、その内容をオペレータに知らせておくことで回答に要する時間の削減を容易にすることができる。

【図面の簡単な説明】

【図１】 自然言語から構文木を作る過程を示す図である。

【図２】 パターンについて示す図である。

【図３】 本発明の全体構成を示す図である。

【図４】 本発明の処理のフローチャートである。

【図５】 言語解析装置の詳細を示す図である。

【図６】 パターン抽出装置を示す図である。

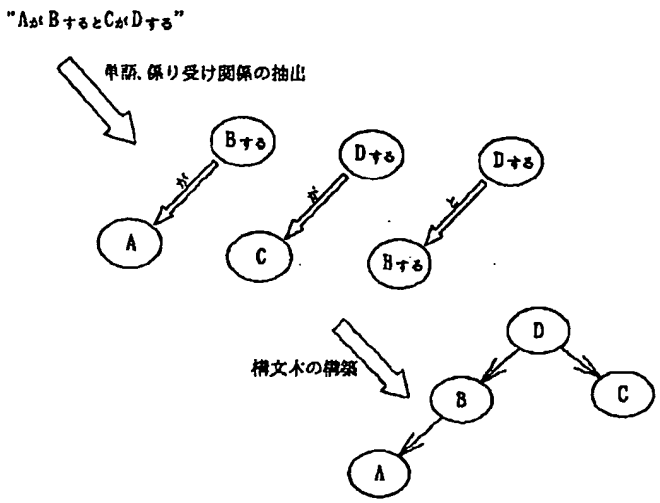
【図７】 抽出された４つの語からなる組（パターン）を示す図である。

【図８】 抽出された６つの語からなる組（パターン）

13

を示す図である。

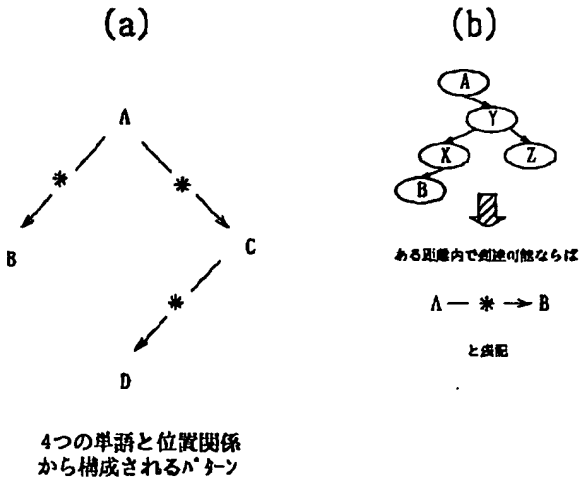
【図１】



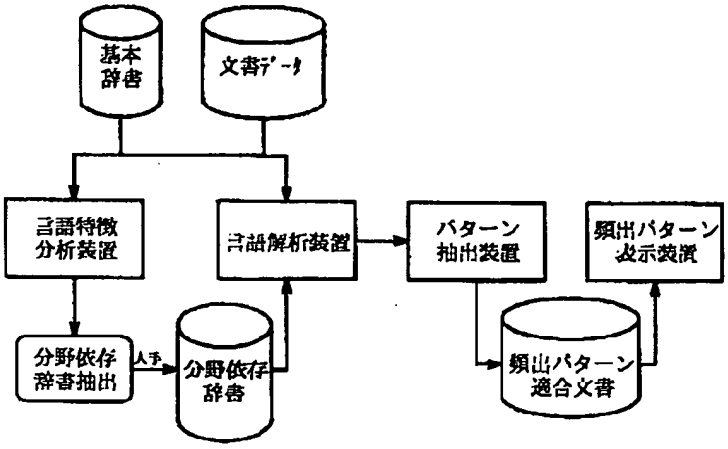
14

【図９】　パターンの例を示す図である。

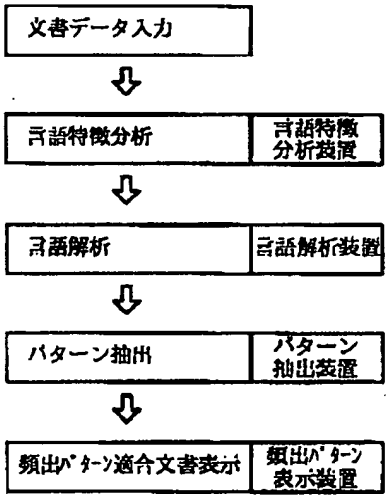
【図２】



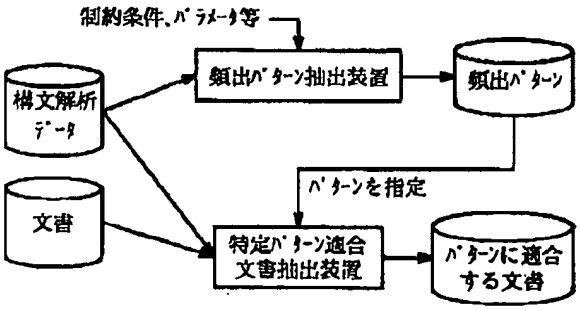
【図３】



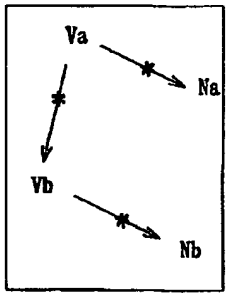
【図４】



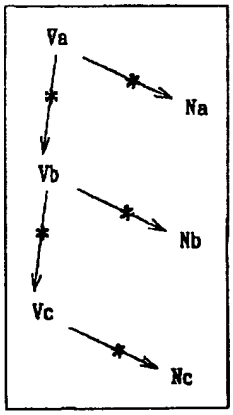
【図６】



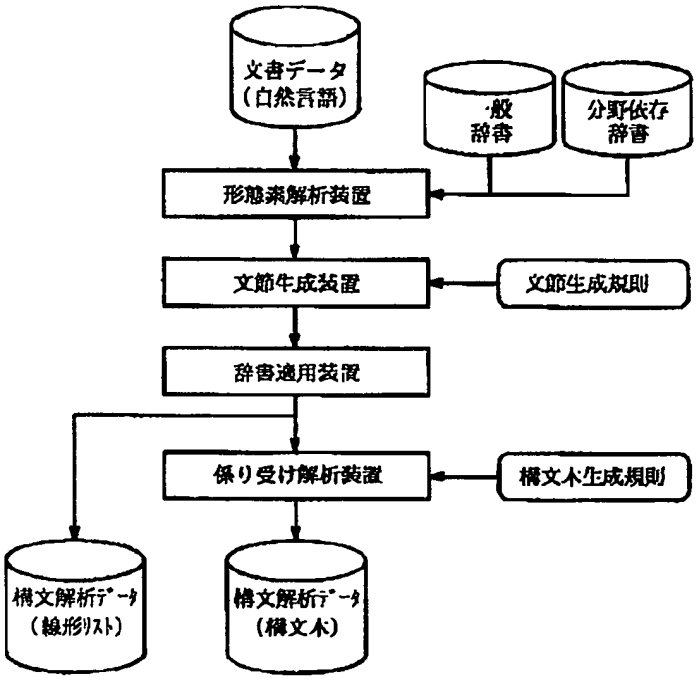
【図７】



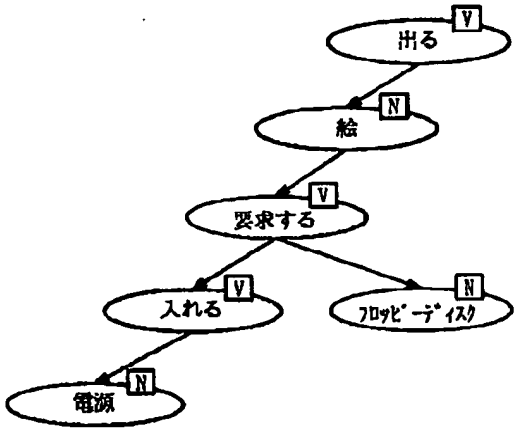
【図８】



【図５】



【図９】



フロントページの続き

(51)Int.Cl.[7]	識別記号	F I	ﾃｰﾏｺｰﾄﾞ(参考)
		Ｇ０６Ｆ　15/401	３３０Ｚ

(72)発明者　松澤　裕史
神奈川県大和市下鶴間1623番地14　日本アイ・ビー・エム株式会社　東京基礎研究所内

(72)発明者　福田　剛志
神奈川県大和市下鶴間1623番地14　日本アイ・ビー・エム株式会社　東京基礎研究所内

(72)発明者　那須川　哲哉
神奈川県大和市下鶴間1623番地14　日本アイ・ビー・エム株式会社　東京基礎研究所内

(72)発明者　長野　徹
神奈川県大和市下鶴間1623番地14　日本アイ・ビー・エム株式会社　東京基礎研究所内

(72)発明者　諸橋　正幸
東京都多摩市聖ヶ丘　４丁目１番地１号　多摩大学経営情報学部内

Ｆターム(参考)　5B075 ND03 NK31 NK32 NK43 PP24 PR04 UU40
5B091 AA15 CA02 CA05 CC01 CC02 CC05